J. Acoust. Soc. Jpn. (*E*) **14**, 2 (1993)

Japanese digits recognition by neural networks using vocal tract shapes

Hiroshi Kinugasa, Hiroyuki Kamata, and Yoshihisa Ishida

School of Science and Technology, Meiji University,
1-1-1, Higashi-mita, Tama-ku, Kawasaki, 214 Japan

(*Received 18 February 1992*)

This paper presents a new system for spoken Japanese digits recognition by a neural network using vocal tract shapes. The vocal tract shape is a suitable parameter for synthesis or recognition. The vocal tract shapes are used for the neural network as input data. We first propose a simple method by which the vocal tract shape is directly estimated from speech waves. A three-layered neural network is used in our recognition system. The network learning algorithms utilized here are conjugate gradient (CG) algorithm and backpropagation (BP) algorithm. Finally, we show the recognition results to prove the effectiveness of our method, and we show that the CG algorithm has several advantages compared to the BP algorithm.

Keywords: Speech recognition, Vocal tract area, Adaptive inverse filter, Neural network, Conjugate gradient method

PACS number: 43. 72. Ne

1. INTRODUCTION

A vocal tract shape is a suitable parameter for speech synthesis or recognition. However, we must eliminate the influence of the sound source and radiation characteristics for an estimate of the vocal tract shapes from speech waves directly. The estimate method proposed by Nakajima *et al.*[1] utilizes the adaptive inverse filter for discrete time signals to eliminate this influence. In this paper, we propose a simple method that the effect of the adaptive inverse filter is given to autocorrelation coefficients directly. We utilize the adaptive inverse filter in order to eliminate the influence of the sound source and radiation characteristics. Our method can reduce the computation time by about 30% in comparison with Nakajima's method.

Multilayer perceptrons have changed to popular algorithm since the development of the BP learning algorithm proposed by Rumelhart *et al.*[2] The BP algorithm attempts to minimize the least squared error objective function, defined by the differences between the actual network outputs and the desired outputs. This paper presents a neural network learning algorithm, which is superior to the conventional BP algorithm in training multilayer networks. The network learning algorithm presented herein is based on the CG method.[8] The algorithm updates the input weights to each neuron in an efficient parallel way, similar to that used by the conventional BP learning algorithm.

This paper is organized as follows. Section 2 gives a simple estimate method of vocal tract shapes that the effect of the adaptive inverse filter is given to autocorrelation coefficients directly. Section 3 presents the network learning algorithm, which is based on the CG method. Section 4 shows the performance of the recognition experiments for 10 Japanese digit patterns using the proposed method.

2. ESTIMATE OF VOCAL TRACT SHAPE

From the speech production mechanism, the total system function $S(z)$ for the speech production system is represented by the product of the system

functions for source generation $G(z)$, vocal tract resonance $V(z)$ and radiation $R(z)$:

$$S(z)=G(z)V(z)R(z) \quad (1)$$

In order to obtain $V(z)$ directly from speech waves, we must eliminate the influence of $G(z)$ and $R(z)$. This operation is called an adaptive inverse filtering. The adaptive inverse filter has been developed by Nakajima *et al.*[1] Figure 1 shows the architecture of the adaptive inverse filter. In the figure, Af_1 to Af_4 and Af_6 are filters that revise the slope in the spectral envelope, which is due to the sound source and radiation characteristics. Af_5 and Af_7 are determined by an experiment, in order to compensate for a curve in the center of the spectral envelope.

The transfer functions of these filters are described as follows:

$$Af_i(z)=1-\epsilon_i z^{-r} \quad (2)$$

The coefficients ϵ_i of these filters are obtained as follows:

$$\epsilon_i=\frac{R(r)}{R(0)} \quad (3)$$

where $r=1$ for $i=1$ to 4, 6
$r=2$ for $i=5$
$r=3$ for $i=7$

$R(\ )$ is the autocorrelation coefficient for the input signal to each filter.

Reflection coefficients k_i at the output of the last stage filter Af_7 can be obtained by using the Levinson-Durbin algorithm. With these reflection coefficients, the vocal tract shape becomes

$$A_{i+1}=\frac{1-k_i}{1+k_i}A_i \quad (4)$$

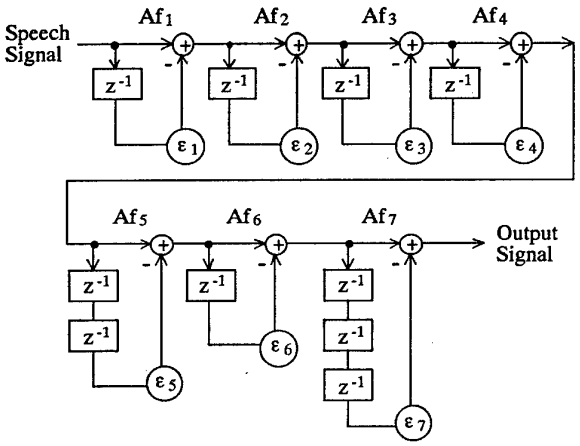


Fig. 1 The architecture of the adaptive inverse filter.

Now, let us assume that we apply the adaptive inverse filter mentioned above to the autocorrelation coefficients.[4] When the power spectrum of the input sequence of the filter Af_i is written as $P_i(m)$, the corresponding autocorrelation coefficient $R_i(k)$ can be obtained from the inverse Fourier transformation of $P_i(m)$ as

$$R_i(k)=\frac{1}{N}\sum_{m=0}^{N-1}P_i(m)\exp\left(j\frac{2\pi mk}{N}\right) \quad (5)$$

Similarly, when the power spectrum of the filter Af_i is written as $F_i(m)$, the corresponding autocorrelation coefficient $Q_i(k)$ can be described by the equation:

$$Q_i(k)=\frac{1}{N}\sum_{m=0}^{N-1}F_i(m)\exp\left(j\frac{2\pi mk}{N}\right) \quad (6)$$

Using the convolution sum, we have

$$R_{i+1}(k)=R_i(k)*Q_i(k) \quad (7)$$

where $*$ is used to denote the convolution sum, and $R_{i+1}(k)$ is the autocorrelation coefficient of the output sequence.

On the other hand, the power spectrum $F_i(m)$ of the filter is written as follows:

$$F_i(m)=|Af_i(z)|^2$$
$$=1+\epsilon_i^2-2\epsilon_i\cos\left(\frac{2\pi rm}{N}\right) \quad \text{at } z=\exp\left(j\frac{2\pi m}{N}\right) \quad (8)$$

The corresponding autocorrelation coefficient $Q_i(k)$ becomes

$$\begin{aligned} Q_i(k)&=1+\epsilon_i^2 && \text{for } k=0\\ &=-\epsilon_i && \text{for } k=\pm r\\ &=0 && \text{for } k\neq 0,\ \pm r \end{aligned} \quad (9)$$

Replacing $Q_i(k)$ in Eq. (7) by its value as given by Eq. (9), the autocorrelation coefficients can be written in the form of simple recursive equation as follows:

$$R_{i+1}(k)=(1+\epsilon_i^2)R_i(k)-\epsilon_i R_i(|k-r|)-\epsilon_i R_i(k+r) \quad (10)$$

Figure 2 shows the architecture for calculating autocorrelation coefficients by Eq. (10).

We can calculate reflection coefficients and the vocal tract shape from the autocorrelation coefficients using the Levinson-Durbin algorithm and Eq. (4). Figure 3 illustrates the vocal tract shapes estimated by Nakajima's method and the proposed method.

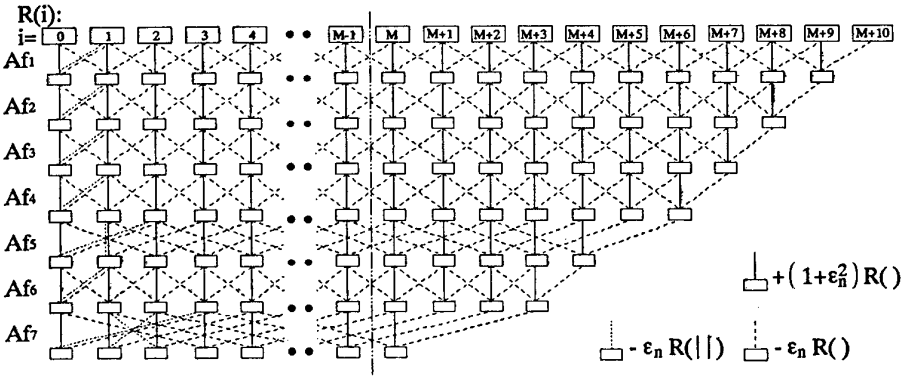


Fig. 2 The architecture for calculating autocorrelation coefficients by Eq. (10).

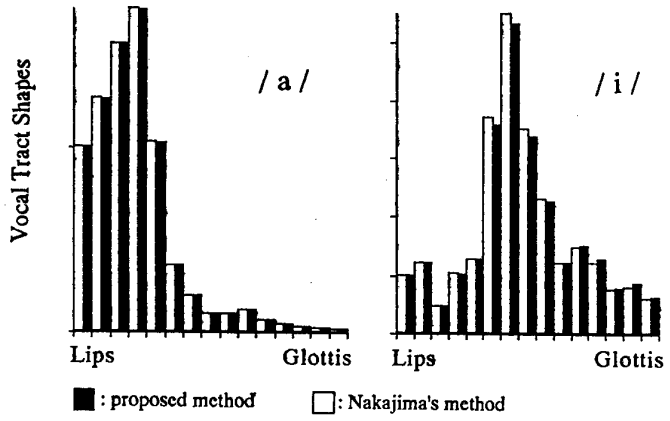


Fig. 3 The vocal tract shapes estimated by Nakajima's method and the proposed method.

3. NETWORK LEARNING ALGORITHM[3]

The learning procedure at the k-th iteration requires to compare the output vector $\boldsymbol{Y}_i$ of the network due to an input pattern vector with a target vector $\boldsymbol{T}_i$.

The resulting error is

$$f_i(\boldsymbol{W})=\frac{1}{2}||\boldsymbol{Y}_i-\boldsymbol{T}_i||^2 \tag{11}$$

where $\boldsymbol{W}$ is the weight vector.

We will try to minimize the least squared function made up from the outputs of all the input patterns:

$$\phi(\boldsymbol{W})=\sum_{i=1}^{p} f_i(\boldsymbol{W}) \tag{12}$$

where p is the number of input pattern vectors.

The gradient vector of $\phi(\boldsymbol{W})$ at point $\boldsymbol{W}$ is given by

$$\nabla\phi(\boldsymbol{W})=\sum_{i=1}^{p} \nabla f_i(\boldsymbol{W}) \tag{13}$$

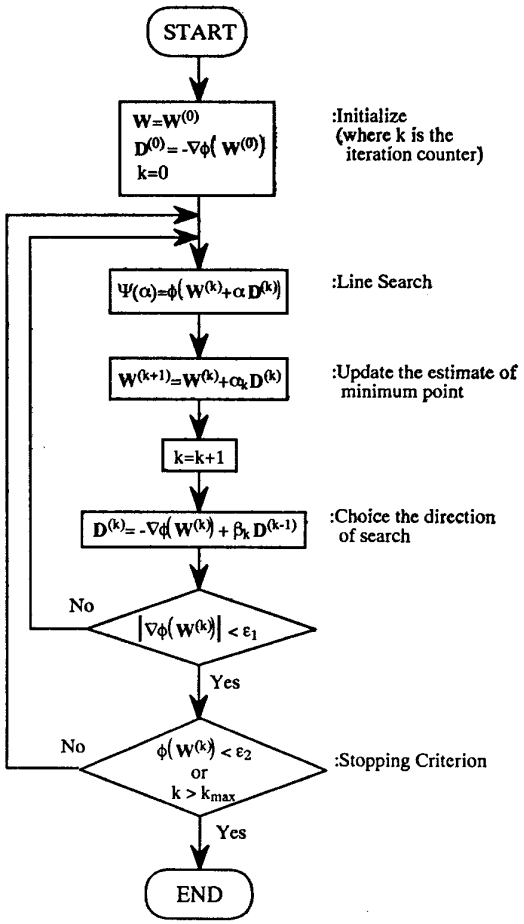


Fig. 4 The conjugate gradient learning algorithm.

We assume the following single variable function:

$$\psi(\alpha)=\phi(\boldsymbol{W}^{(k)}+\alpha\boldsymbol{D}^{(k)}) \tag{14}$$

where $\boldsymbol{W}^{(k)}$ is the weight vector at the k-th iteration and $\boldsymbol{D}^{(k)}$ is the direction based on the conjugate gradient method as follows[5]:

$$\boldsymbol{D}^{(k)}=-\nabla\phi(\boldsymbol{W}^{(k)})+\beta_k\boldsymbol{D}^{(k-1)} \tag{15}$$

where

$$\beta_k=\frac{(\nabla\phi(W^{(k)})-\nabla\phi(W^{(k-1)}))^{\mathrm{T}}\nabla\phi(W^{(k)})}{\nabla\phi(W^{(k-1)})^{\mathrm{T}}\nabla\phi(W^{(k-1)})} \quad (16)$$

Then we can update the weight vector as follows:

$$W^{(k+1)}=W^{(k)}+\alpha_k D^{(k)} \quad (17)$$

where α_k is the value of α that will sufficiently decrease the function value of $\psi(\alpha)$.

It should be noted that the CG algorithm can be considered as the BP algorithm with adjustable coefficients λ and μ, $\lambda_k=\alpha_k$, $\mu_k=\beta_k$, which are the learning coefficient and the momentum coefficient for the BP algorithm, respectively.

The flow chart of the conjugate gradient learning algorithm is shown in Fig. 4.

4. EXPERIMENTAL RESULTS

4.1 The Speech Database

Our speech experiments are performed with a database of 10 Japanese digits (*i.e.*, /ZERO/, /ICHI/, /NI/, /SAN/, /YON/, /GO/, /ROKU/, /NANA/, /HACHI/ and /KYU/. These are written in English respectively as follows: "ZERO," "ONE," "TWO," "THREE," "FOUR," "FIVE," "SIX," "SEVEN," "EIGHT" and "NINE"). The speech data is collected from 8 male speakers. 10 spoken digits are uttered 3 times by each speaker (30 utterances per each speaker). Three speakers utter training words once and test words twice. The remaining five speakers utter the same words 3 times. Thus, the total number of utterances for the training is 30. The total number of speaker-dependent experiments is 90 (including the training patterns). The remaining 150 utterances are used for speaker-independent experiments.

4.2 Preliminary Experiments

At the present, pattern matching is one of the most reliable recognition methods. We first ran some preliminary experiments on the task of recognizing 10 spoken digits using the pattern matching method. A block diagram of speech processing system used in this preliminary experiment is given in Fig. 5. Speech signals are converted into 12-bit samples at a rate of 10 kHz. The beginning and end points of the actual sample utterance are determined by manual inspection. The speech signal is devided into some frames and each frame is multiplied by a Hamming window. The length of each frame is 25.6 ms and successive frames overlap by 10 ms. LPC analysis is carried out and then a vocal tract shape is calculated for each frame. Furthermore, the speech signal in each frame is processed through a High Pass Filter (H.P.F.) with cut-off frequency 2.5 kHz, and the maximum absolute value of the processed speech is divided by the counterpart in the original wave form. Similarly, a Low Pass Filter (L.P.F.) with cut-off frequency 500 Hz is applied. The length of the input speech data is normalized to 16 segments using the linear time warping.

For these speech signals, we carried out the iso-

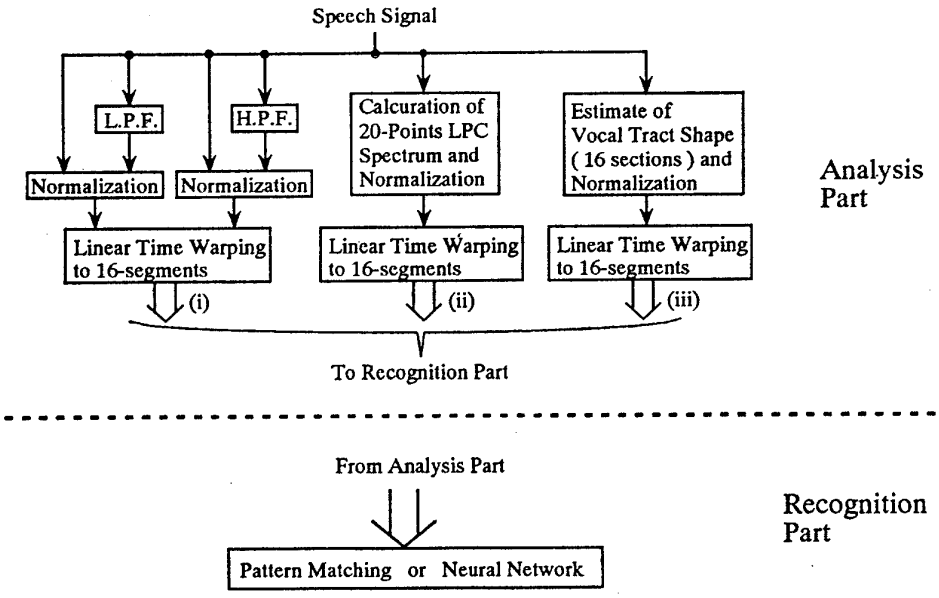


Fig. 5 Processing steps from the speech signal to a standardized word template.

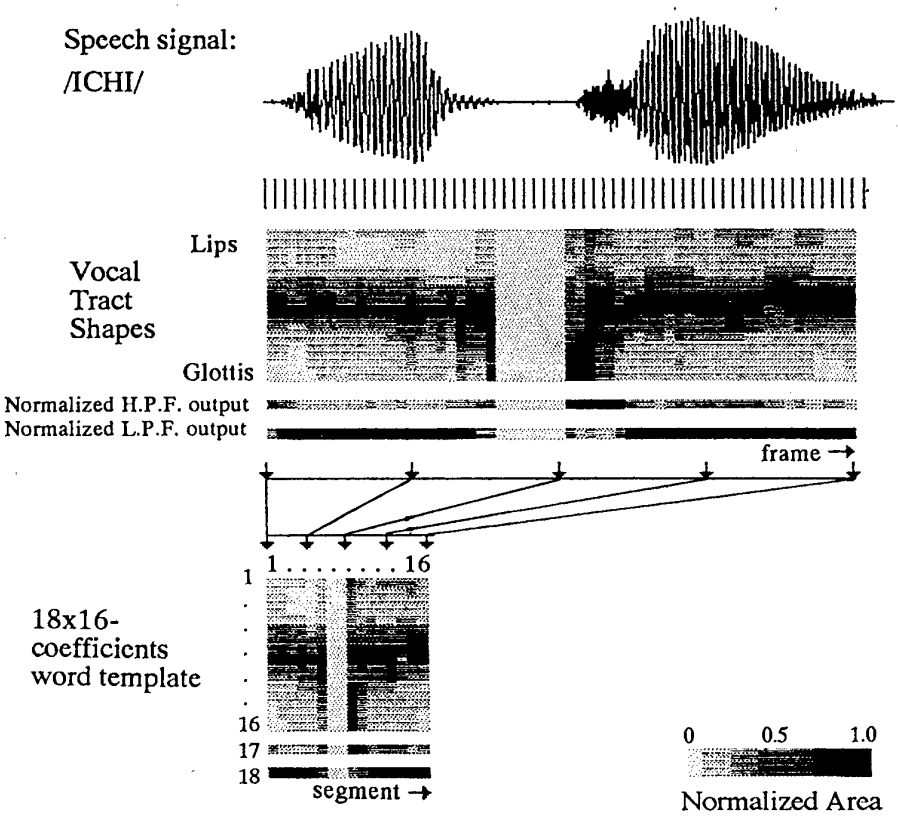


Fig. 6 The processing steps for the word/ICHI/.

lated spoken digits recognition using the pattern matching method. The speaker-independent experimental results are summarized as follows:

(1) When the low-frequency portion and high-frequency portion of the power spectrum were used as the feature parameters, a mean recognition rate was 20.7%. ((i) in Fig. 5)

(2) When the 20-points LPC spectra were used, a mean recognition rate was 86.7%. ((ii) in Fig. 5)

(3) When the 16-sections vocal tract shapes were used, a mean recognition rate was 94.7%. ((iii) in Fig. 5)

In these experiments, we used the average values of test utterances by 3 speakers as the reference templates. The recognition using the vocal tract shapes has a higher score than that using the LPC spectra. We consider that this is because the influence by the moving of the place of articulation appears more clearly on the vocal tract shape.[1] These results suggest that the vocal tract shape will be useful for improving speech recognition.

Based on the results described above, we decided to use the vocal tract shape and low/high frequency components as the feature parameters for the digits recognition. As an example, we show the processed spoken digit /ICHI/ in Fig. 6.

4.3 The Neural Network

Pattern matching has excellent classification performance. However, it needs a laborious task to construct reference templates for each speaker. So we next implemented a digits recognition system using a neural network. The advantages of using a neural network are summarized as follows:

(1) It does not need the reference templates for each speaker.

(2) It does not take more computation time to obtain the recognition result compared with pattern matching.

(3) It can easily carry out the additional learning for new speakers.

(4) It can obtain almost the same recognition score as pattern matching (This will be obvious in Section 4.4.).

A three-layered neural network is shown in Fig. 7. The network has 288 (18×16) units in the input layer, 30 hidden units, and 10 units in the output layer. Both the 16-sections vocal tract shape and the low/high-frequency portions of the power spectrum described in 4.2 are applied to the input layer as a

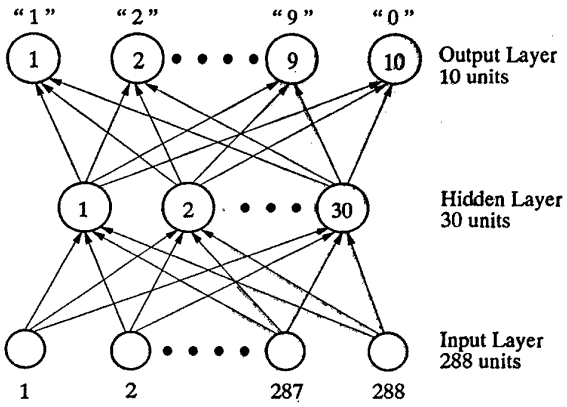


Fig. 7 A three-layered neural network.

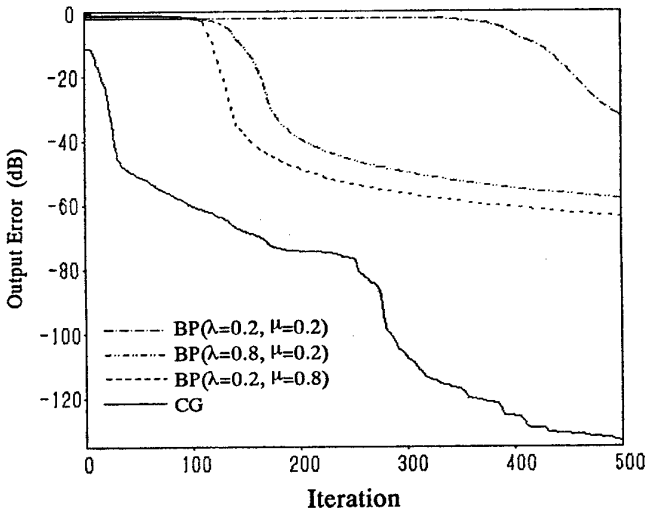


Fig. 8 The convergence of output error with the CG algorithm and the BP algorithm.

set of word templates. We use the low/high frequency components as supplement because, in case of nasal and fricative consonants, the reasonable vocal tract shapes are not occasionally obtained. 10 output units correspond to 10 Japanese digits, respectively.

Figure 8 shows the convergence of output error with the CG algorithm and one with the BP algorithm for three sets of the learning coefficient λ and the momentum coefficient μ.

As well known, the BP algorithm is an effective network learning algorithm and can be easily installed on the computer. However, this algorithm often exhibits oscillatory behavior and even diverges when the values of coefficients λ and μ have not been determined adequately. Furthermore, it is generally difficult to find the most suitable values of λ and μ. In our experiments of the BP algorithm, the values ($\lambda=0.2$, $\mu=0.8$) produced the best performance. Hereafter these values are used for the experiments.

On the other hand, the CG algorithm described in this paper can automatically adjust the learning parameters so that the learning procedure is effective. From Fig. 8, we see that the convergence for the BP algorithm depends on the values of the coefficients λ and μ, and the superiority for the convergence of the CG algorithm over the BP algorithm is apparent. We also confirm this superiority of the CG algorithm for the convergence in the study on system identification.[6]

4.4 Results

In the result tables, the numerator in a parenthesis is the number of correct answer and the denominator is the total number of spoken digits used for the judgment. Table 1 is the results of the speaker-independent test by neural networks using the CG algorithm and the BP algorithm, where the output error is -40 dB. Then the learning iteration by the CG algorithm is 28 and the one by the BP algorithm is 160. From Table 1, the recognition rates by the CG algorithm and the BP algorithm are 98.6% and 98.0%, respectively. These results clearly show the effectiveness of our method. From the above results, for the number of iteration, the superiority of the CG algorithm over the BP algorithm is apparent. Table 2 shows the results of the speaker-dependent test where the output error is -40 dB. In this table, each of the recognition results is 100%.

Table 3 shows the results of the speaker-independent test where the learning iteration is 200. Then the

Table 1 Recognition results of speaker-independent test in percent. (Where output error is -40 dB.)

Algorithm	0	1	2	3	4	5	6	7	8	9	Total
CG	100	100	100	100	100	92.3	100	92.9	100	100	98.6 (142/144)
BP	100	100	100	100	100	93.3	92.9	93.3	100	100	98.0 (144/147)

Table 2 Recognition results of speaker-dependent test in percent. (Where output error is −40 dB.)

Algorithm	0	1	2	3	4	5	6	7	8	9	Total
CG	100	100	100	100	100	100	100	100	100	100	100 (90/90)
BP	100	100	100	100	100	100	100	100	100	100	100 (90/90)

Table 3 Recognition results of speaker-independent test in percent. (Where learning iteration is 200.)

Algorithm	0	1	2	3	4	5	6	7	8	9	Total
CG	100	100	92.3	100	100	100	92.3	100	100	100	98.5 (131/133)
BP	100	100	100	100	100	93.3	100	100	100	100	99.3 (143/144)

Table 4 Recognition results of speaker-dependent test in percent. (Where learning iteration is 200.)

Algorithm	0	1	2	3	4	5	6	7	8	9	Total
CG	100	100	100	100	100	100	100	100	100	100	100 (90/90)
BP	100	100	100	100	100	100	100	100	100	100	100 (90/90)

output error by the CG algorithm is −74.4 dB and the one by the BP algorithm is −48.5 dB. As shown in Table 3, the total recognition rates by the CG algorithm and the BP algorithm are 98.5% and 99.3%, respectively. Table 4 shows the results of the speaker-dependent test where the learning iteration is 200. As shown in Table 4, each of the recognition results is 100%.

By the way, according to the recognition experiment using both the pattern matching method and the same word templates as those described in section 4.3, a mean recognition rate was 98.7%. We see that the recognition system using a neural network has almost the same recognition rate as that using the pattern matching method.

5. CONCLUSION

In this paper, we have proposed a new method for the spoken Japanese digits recognition by a neural network using vocal tract shapes. The experimental results have shown that the vocal tract shape is useful for improving of spoken digits recognition. We have also presented a neural network learning algorithm, which is based on the CG method. As the results, we have shown that the convergence of the CG algorithm is superior to that of the conventional BP algorithm.

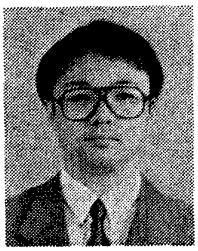
ACKNOWLEDGMENTS

The authors greatly wish to thank Prof. Y. Ogawa of Meiji University and Prof. C. Charalambous of Kuwait University for their helpful advices. The authors also wish to thank the members of the Instrument and Control Laboratory of Meiji University for their valuable discussions.

REFERENCES

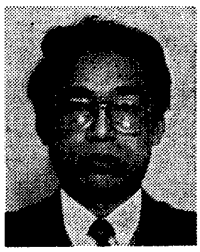
1) T. Nakajima, T. Suzuki, H. Ohmura, S. Ishizaki, and K. Tanaka, "Estimation of vocal tract area function by adaptive deconvolution and adaptive speech analysis system," J. Acoust. Soc. Jpn. (J) **34**, 157–166 (1978) (in Japanese).
2) D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing*, Vol. 1 (MIT Press, Cambridge MA, 1986).
3) Y. Ishida, K. Endo, H. Kamata, and C. Charalambous, "Vowel recognition by a neural network using vocal tract area function," Proc. IJCNN 91 Singapore **3**, 2144–2149 (1991).
4) H. Kamata, Y. Ishida, and Y. Ogawa, "High speed estimation of vocal tract area function using digital signal processor," Trans. IEE Jpn. **110-D-7**, 773–780 (1990) (in Japanese).
5) E. Polak, *Computational Methods in Optimization: A Unified Approach* (Academic Press, New York,

1971).
6) K. Endo and Y. Ishida, "System identification using neural networks," Proc. 2nd IFAC Workshop on AARTC, 129–134 (1992).



Hiroshi Kinugasa was born in Tokyo, Japan, on January 21, 1968. He received the B.E. degree in electrical engineering from Meiji University, Kawasaki, Japan, in 1991. He is currently working toward M.E. degree at the Department of Electrical Engineering, Meiji University. His current research interests are in the area of neural networks and speech recognition. He is a member of the IEICE of Japan.

Hiroyuki Kamata was born in Kawasaki, Japan, on December 1, 1959. He received the B. E., M. E., and Dr. Eng. degrees in electrical engineering from Meiji University, Kawasaki, Japan, in 1982, 1984 and 1987, respectively. In 1987, he joined the Department of Electronics and Communication, Meiji University, as a Research Assistant. He is currently a Lecturer at the Department of Electronics and Communication, Meiji University. His current research interests are in the area of digital signal processing, speech analysis, synthesis and recognition. He is a member of the Acoustical Society of Japan, IPS of Japan and IEICE of Japan.



Yoshihisa Ishida was born in Tokyo, Japan, on February 24, 1947. He received the B.E., M.E. and Dr. Eng. degrees in electrical engineering from Meiji University, Kawasaki, Japan, in 1970, 1972, and 1978, respectively. In 1975 he joined the Department of Electrical Engineering, Meiji University, as a Research Assistant and became a Lecturer and an Associate professor in 1978 and 1981, respectively. He is currently a Professor at the Department of Electronics and Communication, Meiji University. His current research interests are in the area of digital signal processing, speech analysis and recognition, and digital control. He is a member of the Acoustical Society of Japan, the IEEE and the IEICE of Japan.